DOCUMENT RESUME

ED 223 687 TM 820 831

AUTHOR Masters, Geoff N.; Wright, Benjamin D.
TITLE Some Fit Issues in Rating Scale Analysis.
PUB DATE Mar 82
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *Goodness of Fit; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Maximum Likelihood Statistics; Probability; *Rating Scales; Response Style (Tests); Scoring; Test Items
IDENTIFIERS Partial Credit Model; *Rasch Model

ABSTRACT

The analysis of fit of data to a measurement model for graded responses is described. The model is an extension of Rasch's dichotomous model to formats which provide more than two levels of response to items. The model contains one parameter for each person and one parameter for each "step" in an item. A dichotomously-scored item provides only one step from fail to pass so only one item parameter is estimated. An item with three ordered performance levels (0, 1, and 2), provides two steps. For this item two parameters are estimated: the first governs the model probability of scoring 1 rather than 0, and the second governs the probability of scoring 2 rather than 1. The model probability of completing the k'th step in an item is described by a simple logistic ogive of slope one and is governed by the person's ability and the difficulties of the step. This "Partial Credit" model's procedures for estimating the parameters are unconditional maximum likelihood. Both the effects of individual differences in response style and variations in the fit of items are examined, including items with large positive or large negative standardized mean squares. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223687

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Masters

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SOME FIT ISSUES IN RATING SCALE ANALYSIS

Geoff N. Masters

Education Department of Western Australia

and

Benjamin D. Wright

University of Chicago

TM 820 831

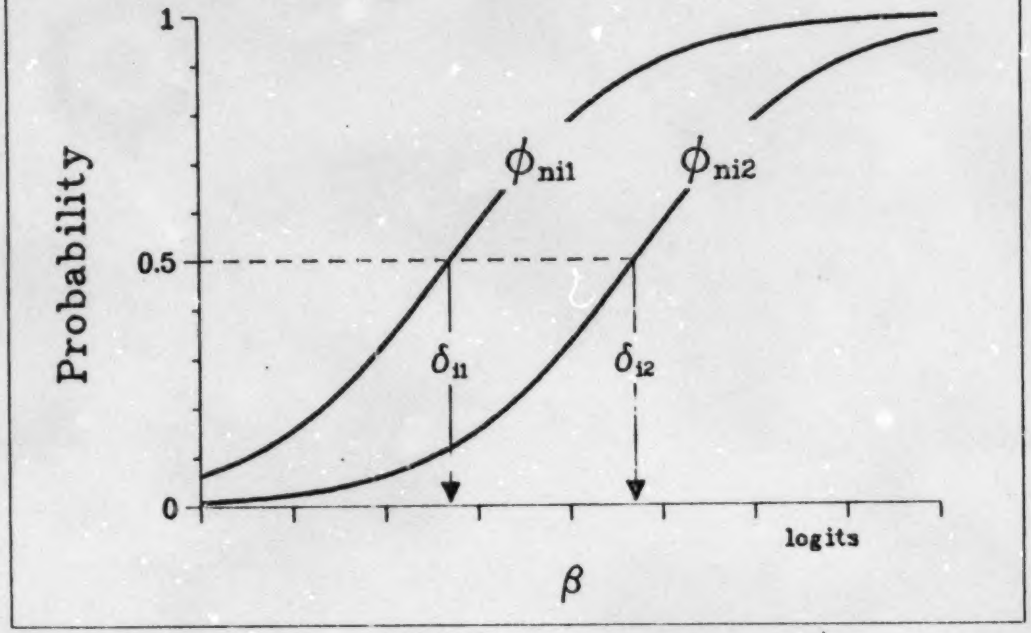
Paper presented at a symposium on analyzing and understanding the fit of data to Rasch models, Annual Meeting of the American Educational Research Association, New York, March, 1982.

This paper describes some of our efforts and experience in analyzing the fit of data to a measurement model for graded responses. The model we use is an extension of Rasch's Dichotomous model to formats which provide more than two levels of response to items. We have used it to analyze a wide variety of attitude questionnaires, morale inventories, performance ratings and achievement tests. You can study some of these analyses in our book Rating Scale Analysis which is now available from MESA Press.

The model is developed very simply. It contains one parameter for each person and one parameter for each "step" in an item. A dichotomously-scored item provides only ONE step from fail to pass, and so for a dichotomously-scored item, only ONE item parameter is estimated. This is the usual item difficulty δ_i. An item with three ordered performance levels, 0, 1 and 2, provides TWO steps. For this item two parameters are estimated. The first, δ_{i1}, governs the model probability of scoring 1 rather than 0. The second, δ_{i2}, governs the model probability of scoring 2 rather than 1. This is illustrated in Figure 1 of the handout.

Figure 1 shows the operating curves for a two-step item i. These are logistic ogives of the same slope which differ only in their location on the variable. In this case, δ_{i2} is to the right of δ_{i1} meaning that the second step is harder to complete than the first. But this order is not necessary. The second step in an item may be easier than the first, even though it can only be attempted after the harder first step has been completed. In general, the model probability of a person completing the k'th step in an item is described by a simple logistic ogive of slope one and is governed only by the person's ability β_n and the difficulty δ_{ik} of that step.

FIGURE *1*
ITEM OPERATING CURVES FOR
A TWO-STEP ITEM *i*



Partial Credit Model

$$\pi_{nix} = \frac{\exp\sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(\beta_n - \delta_{ij})} \qquad x = 0, 1, \ldots, m_i$$

From: Wright and Masters, (1982)

We call this extension of Rasch's Dichotomous model the "Partial Credit" model. It is also algebraically straightforward. The model appears directly below Figure 1. π_{nix} is the model probability of person n scoring x on item i. This probability is governed by the person's ability β_n and the difficulties of the m_i steps in item i. The numerator contains only the difficulties of the x steps which must be taken to make a score of x on item i. The denominator is the sum of all possible numerators.

In our book we describe four procedures for estimating the parameters in this model-- a conditional maximum likelihood procedure, an unconditional maximum likelihood procedure, a pairwise procedure and a normal approximation procedure. In our analyses we have used only the unconditional procedure.

Table 1 shows the details of our item fit statistics. The observed response x_{ni} takes values between 0 and m_i, the number of steps in item i. The item fit statistics we report routinely are the weighted mean square and the standardized weighted mean square at the bottom of the left column of Table 1. The weighted mean square differs from the unweighted mean square above it in that it is less influenced by the surprising responses of persons for whom the item is too easy or too hard. When data fit the model, the standardized weighted mean sqaure is approximately normally distributed with a mean near zero and a standard deviation near one. To analyze the fit of persons' response records to the model we calculate these same fit statistics, but now summing over items for each person rather than over persons for each item.

TABLE 1

CALCULATING ITEM FIT STATISTICS

Observed Response

x_{ni}

Expected Mean of x_{ni}

$$E_{ni} = \sum_{k=0}^{m_i} k\pi_{nik}$$

$$\pi_{nik} = \exp\sum_{j=0}^{k}(\beta_n - \delta_{ij}) / \Psi_{ni}$$

Variance of x_{ni}

$$W_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^2 \pi_{nik}$$

$$\Psi_{ni} = \sum_{k=0}^{m_i} \exp\sum_{j=0}^{k}(\beta_n - \delta_{ij})$$

Kurtosis of x_{ni}

$$C_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^4 \pi_{nik}$$

	Expectation	Variance
Score Residual $y_{ni} = X_{ni} - E_{ni}$	0	W_{ni}
Standardized Residual $z_{ni} = y_{ni}/W_{ni}^{1/2}$	0	1
Score Residual Squared $y_{ni}^2 = W_{ni}z_{ni}^2$	W_{ni}	$C_{ni} - W_{ni}^2$
Standardized Residual Squared z_{ni}^2	1	$(C_{ni}/W_{ni}^2) - 1$
Unweighted Mean Square $u_i = \sum_{n=1}^{N} z_{ni}^2/N$	1	$\sum_{n}^{N}(C_{ni}/W_{ni}^2)/N^2 - 1/N$
Weighted Mean Square $v_i = \sum_{n}^{N} W_{ni}z_{ni}^2 / \sum_{n}^{N} W_{ni}$ $= \sum_{n}^{N} y_{ni}^2 / \sum_{n}^{N} W_{ni}$	1	$q_i^2 = \sum_{n}^{N}(C_{ni} - W_{ni}^2) / (\sum_{n}^{N} W_{ni})^2$
Standardized Weighted Mean Square $t_i = (v_i^{1/3} - 1)(3/q_i) + (q_i/3)$	0	1

From: Wright and Masters (1982)

Our efforts at analyzing the fit of data to this model have drawn our attention to several new fit issues which arise when responses are scored in more than two ordered response categories. Our analyses have also shed light on some continuing issues in the analysis of fit to Rasch's Dichotomous model.

One new fit issue which arises is that of individual differences in response style. It has long been recognised that differences in response style can interfere with valid measurement. When attitudes are measured, for example, it is sometimes possible to find respondents with unusual tendencies to give only extreme responses like STRONGLY disagree and STRONGLY agree. Others display unusual tendencies to give only "neutral" responses. And, when performances are rated, some judges display unusual tendencies to use only extreme rating categories like POOR and EXCELLENT, while others display unusual tendencies to rate all performances as mediocre.

The observation of such individual differences among raters has led some researchers to recommend that the rating format be abandoned in favor of a dichotomous scoring of responses. Using our analysis we have been able to identify differences in response style and to assess their impact on measurement. This is illustrated in Table 2.

Table 2 shows the responses of two children to a twenty-five item questionnaire constructed to measure liking-for-science. Responses were recorded in three categories, DISLIKE (scored 0), NOT SURE/DON'T CARE (scored 1) and LIKE (scored 2). The twenty-five items are ordered from those easiest to like on the left to those hardest to like on the right. The responses of these two children are relatively orderly. Both like the easy-to-like activities on the far left and dislike the hard-to-like activities on the far right. Both made a score of 28 on the questionnaire. But, the ways in which they

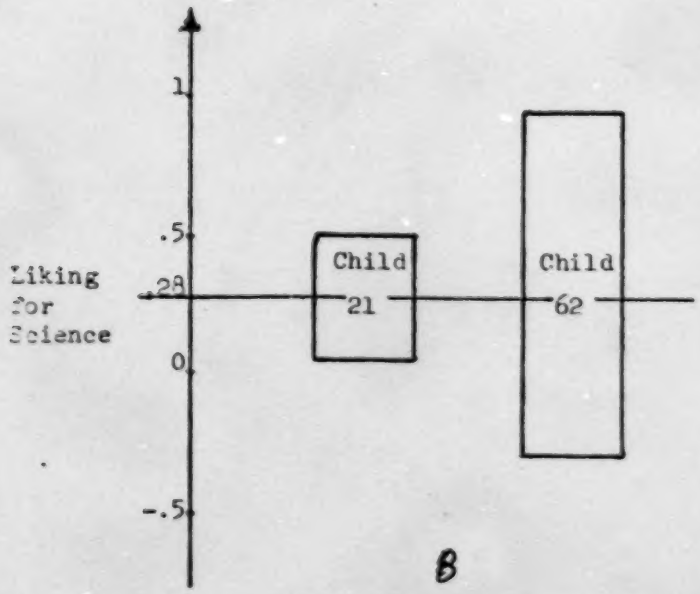
TABLE 2

RESPONSES OF CHILDREN 21 AND 62 TO
THE LIKING-FOR-SCIENCE SCALE

Child	Easy to Like responses			Hard to Like
21	2 2 2 2 2 1 2 1	1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 0 0 0	
62	2 2 2 2 2 2 2 2	2 2 0 0 2 0 2 2 0 0 2 0	0 0 0 0 0	
	(2)	(1)	(0)	

Child	Score	Measure	Error	MnSq v_i	Error q_i	Fit t_i
21	28	.28	.33	.29	.3	-3.94
62	28	.28	.33	1.52	.3	1.87

Child	Min Score	Max Score	Measure		
			Min	Max	Diff
21	26	30	.07	.50	.43
62	22	34	-.35	.95	1.30



made this score of 28 are quite different. Child 21 gave the neutral response to 16 of the 25 items. Child 62 did not give any neutral responses.

In our analysis of these data Child 21 was brought to our attention by his fit statistic of -3.94. His surprising tendency to avoid extreme responses and to favor the middle we refer to as "conservatism". We have observed this tendency on a number of occasions in a variety of instruments, including attitude questionnaires, morale inventories and performance ratings. It results in a weighted mean square less than its expected value of one (that is, a standardized weighted mean square which is large and negative).

The responses of these two children are divided into three regions. According to the model, a child with a score of 28 will most probably score 2 on the eight items on the left, 0 on the five items on the right, and 1 on the twelve items in the middle of Table 2. However, while a score of 1 is the most probable outcome in the middle region of the table, there is also some probability that a person with a score of 28 will score 0 or 2 on these items. The fact that Child 21 made NO 0's and 2's on these middle twelve items is in itself surprising. Child 21's response record is actually more orderly than our probabilistic model expects.

The fact that Child 62 did not score 1 on any of the twelve items in this middle region is also surprising. We call this behavior "extremism". We have found that surprising tendencies to avoid middle categories and to give extreme responses produce moderate to large positive fit statistics.

One approach to interpreting the impact of these differences in response style on the measure we report for these two children is to calculate a "region of uncertainty" about each child's measure. We do this by calculating a maximum possible score and a minimum possible score for each child.

The maximum possible score is the score the child would have made had his surprisingly low responses been those expected of a child with a score of 28. His minimum score is the score he would have made had his surprisingly high responses been the lower ones expected. We have calculated these minimum and maximum possible scores for children 21 and 62 and transformed them to their logit equivalents. These are shown at the bottom of Table 2.

We see that the difference between the minimum and maximum scores for Child 21 corresponds to only .43 logits. The same difference for Child 62 corresponds to 1.30 logits. The smaller region of uncertainty for Child 21 results from his very orderly pattern of responses which leaves us in little doubt about his position on this liking-for-science variable. The larger region of uncertainty for Child 62 results from the overlap between the wide range of items he LIKES, and the wide range of items he DISLIKES. This overlap leaves us uncertain about his position on the liking-for-science variable.

Another response style which has concerned researchers is acquiescence, or the tendency to agree with statements regardless of their content. The usual approach to this problem is to write both positively and negatively worded questionnaire items in the hope that the impact of this tendency-to-agree will be "cancelled out" of the person's measure. Our preference is to view this response style as a source of misfit like "guessing", "carelessness", "conservatism" and "extremism", to <u>expose</u> such unmodelled variation among raters and to assess its impact on measurement.

When a questionnaire contains both positively and negatively worded statements we routinely analyze the two sets of items separately to obtain a measure and fit value for each person on each set. These are then compared to see if they support the combining of positively and negatively worded items.

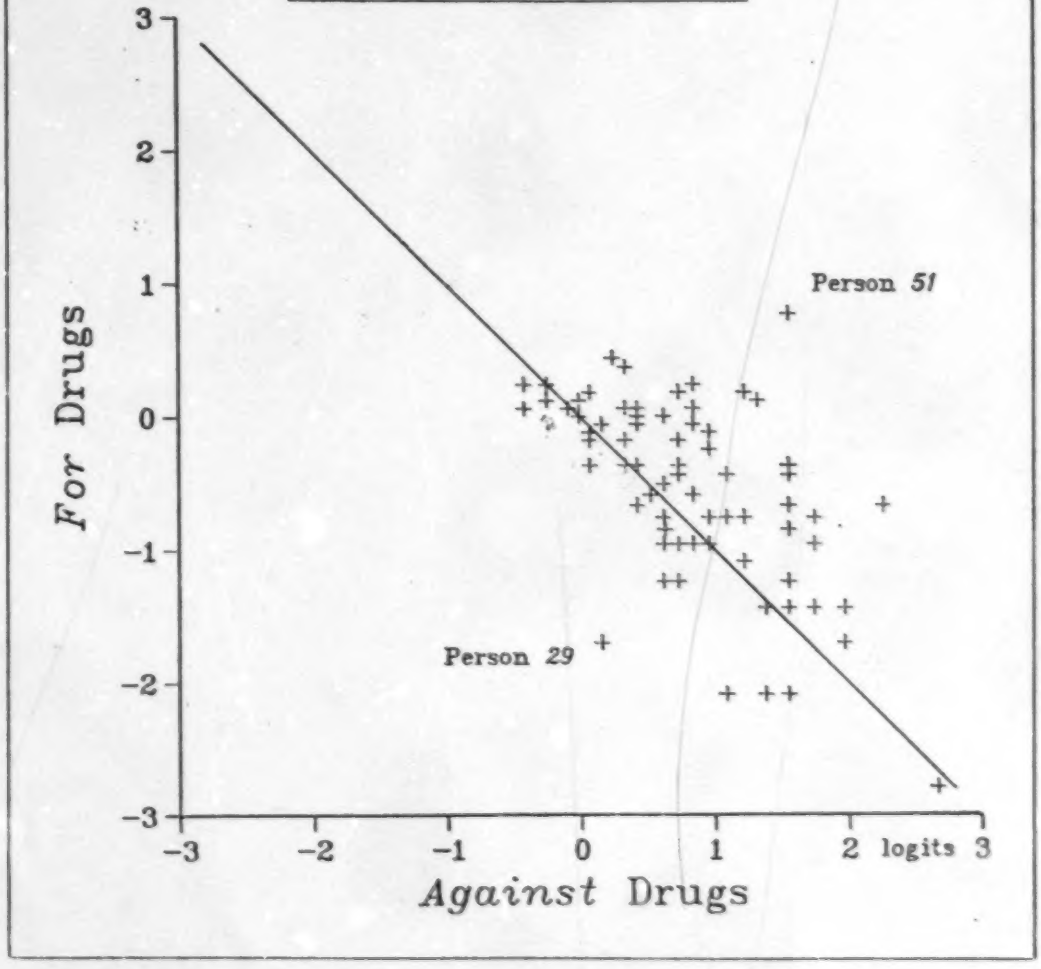
Figure 2 is a plot of attitude estimates for a sample of college students. Each student's measure on ten statements FOR drugs (that is, ten statements favoring drug use) is plotted against his measure on ten statements AGAINST drugs. If these two sets of statements define the same attitude to drugs variable, then students who are most FOR drugs should also be least AGAINST drugs, and those most AGAINST drugs should be least FOR. In other words, we expect points in this plot to lie on or near a line of slope -1.

Two students lie some distance from the inverse identity line in Figure 2. They are persons 29 and 51. As the plot shows, Person 29 is among the students least for drugs, but surprisingly, is also among the students least against drugs! And, Person 51 who is most for drugs, is also among those most against! This plot led us to examine the response records of these two students to see if we could determine why the attitude measures for these two children were inconsistent. These are shown in Table 3.

Responses to the drug statements were recorded in five ordered categories, from Strongly Disagree (scored 0) to Strongly Agree (scored 4). Table 3 shows that Student 29 strongly disagreed with all but two of the FOR items, but also strongly disagreed with all but three of the against items. This student has shown a surprising tendency to <u>disagree</u> with all items, regardless of whether they are for or against drugs. Similarly, student 51 strongly agreed with most FOR items, but also strongly agreed or agreed with four AGAINST items. This student displays a surprising tendency to <u>agree</u> with items of both types.

The responses of these two students are inconsistent and so do not point to a single attitude towards drugs. The usual procedure of adding responses to FOR and AGAINST items in an attempt to "cancel out" such response styles merely hides the problem and does not provide a meaningful attitude measure for either student. The approach we take provides an opportunity to expose such tendencies as threats to valid measurement.

FIGURE 2
ATTITUDE MEASURES FROM *FOR* AND *AGAINST* SCALES



From: Wright and Masters, (1982)

TABLE 3

RESPONSES OF STUDENTS 29 AND 51
TO THE ATTITUDE TO DRUGS QUESTIONNAIRE

Student	Responses FOR	AGAINST
29	3 0 0 0 0 3 0 0 0 0	0 0 0 4 4 0 0 0 0 1
51	4 4 4 4 1 3 2 3 4 4	4 4 4 3 0 1 1 1 0 0

	Measure FOR	Measure AGAINST	Fit FOR	Fit AGAINST
29	-1.55	-0.77	1.50	2.59
51	1.32	-0.12	.93	-.37

	Comparing For and Against: Measure Difference	Error of Diff	Standardized Difference
29	-2.32	.52	-4.4
51	1.20	.47	+2.5

We now consider some issues which have arisen in our attempts to analyze the fit of items to the Partial Credit model. Paralleling our experience with the Dichotomous model we have encountered not only items with large positive standardized mean squares, but also items with large negative standardized mean squares. Two such items are shown in Figures 3 and 4.

The items in Figures 3 and 4 come from a screening test for identifying potential learning problems among prekindergarten children. All items on this test contain three steps. Item 6 in Figure 3 was the item with the most positive misfit in our analysis. Item 5 was the item with the most negative misfit.

Careful inspection of responses to misfitting Item 6 showed that its misfit was due almost entirely to a surprising pattern of failures and successes on its third step. The task in Item 6 is to touch the thumb of each hand to the fingers of that hand in sequence. A child scores 1 if he can touch all fingers on one hand, but not in sequence, and 2 if he touches the fingers of that hand in sequence. The third step is to repeat the task on the other hand.

In Figure 3 we have grouped the 500 children in this sample by ability and calculated the proportion of children in each ability group succeeding on the third step (that is, the proportion of children completing two steps who went on to complete the third). Clearly, more low ability children and fewer high ability children than expected were successful on this third step. This suggests that the third step in Item 6, which requires a degree of ambidextrousness, may not define quite the same variable as the majority of the fine-motor tasks on this test.

For Item 5 in Figure 4 we were able to trace the misfit to the second step

FIGURE 3

OBSERVED PROPORTIONS AND MODEL
PROBABILITIES FOR STEP 3 IN ITEM 6

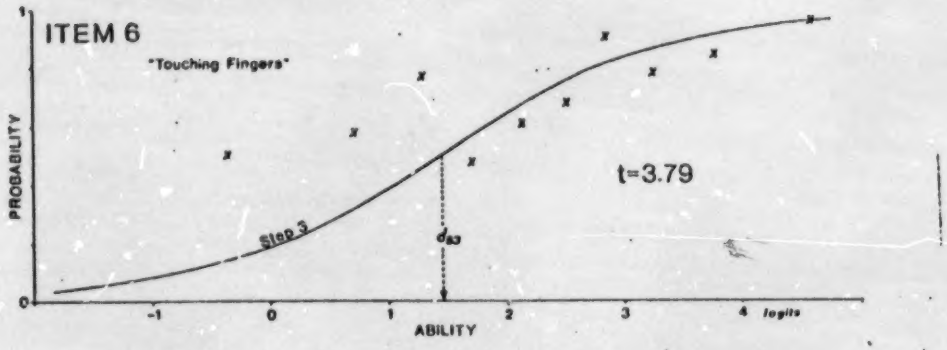
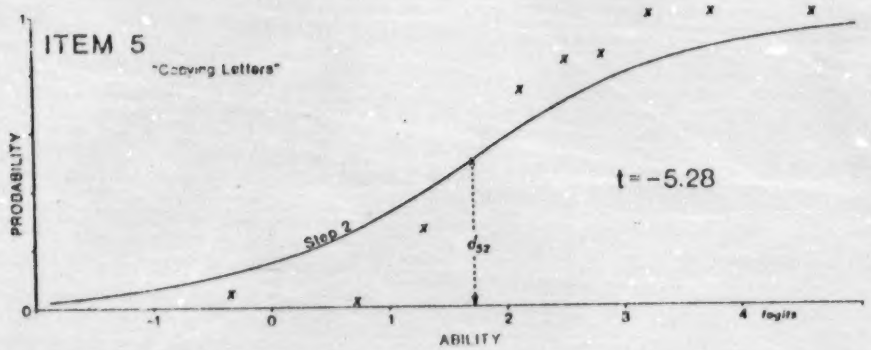


FIGURE 4
OBSERVED PROPORTIONS AND MODEL
PROBABILITIES FOR STEP 2 IN ITEM 5



in the item. To take this second step a child had to draw recogizable copies of letters rather than just scribbling. Fewer low ability children and more high ability children than expected were successful on this step. Dichotomously-scored items frequently show this type of misfit when there is an interaction with instruction. Typically, children who have been exposed to special instruction do better than expected, while thosewho have not do worse.

Interaction with instruction is also a plausible explanation for the misfit of this item. Among children between two and five years of age, copying and drawing are skills which are highly correlated with age and exposure to school. In this sample the older and more able children will probably have had some exposure to classroom settings in which they have had an opportunity to practise copying and drawing. This hypothesis is supported by the observation that the only other item on this test which requires copying or drawing showed an identical pattern of misfit.

What is interesting about Items 5 and 6 is that we were able to trace their misfit to a particular step in each item. When performances on these misfitting steps were studied, the patterns of misfit familiar in the analysis of dichotomously-scored items were found. The opportunity to trace the misfit of an item to a particular step in that item is an advantage of this model, and has obvious implications for item revision. In Item 6, for example, it seems likely that item fit could be improved by revising the third step so that it no longer assesses ambidextrousness. The misfit of Item 5 suggests that in the construction of this screening test it may be preferable to avoid items which assess a skill which some children are likely to have had the opportunity to learn.

Finally, we present some findings from our analysis of a rating scale to assess the professional performance of medical interns. Each item on this scale described an aspect of clinical performance. Senior physicians rated interns on each item using eight rating points from POOR to EXCELLENT. We analyzed data for three successive cohorts of interns (1973, 1974 and 1975).

One item on this instrument showed particularly poor fit to the model. This was item 7. In each year it was by far the worst fitting item with fit values of -9.5, -8.5 and -8.2. The estimated operating curves for the seven steps in this item are shown at the top of Figure 5. These are based on the 1973 results and have been estimated using a special case of the Partial Credit model which assumes the same pattern of operating curves for each item.

Once again, we have grouped persons by ability and calculated the proportions of persons in each group succeeding on each item step. These are shown at the bottom of Figure 5. Again we see that the observed proportions rise much more steeply than the corresponding model probabilities. And, as in Figure 4, this pattern of misfit is associated with a large negative fit value. Now, however, the misfit cannot be attributed to any particular step. <u>All</u> steps in the item are "too discriminating". A consequence of this pattern of misfit is that scores on Item 7 are more highly correlated with total score than the model expects.

The wording of Item 7 suggests a very plausible explanation for its misfit. Item 7 is the last item on the instrument. It reads:

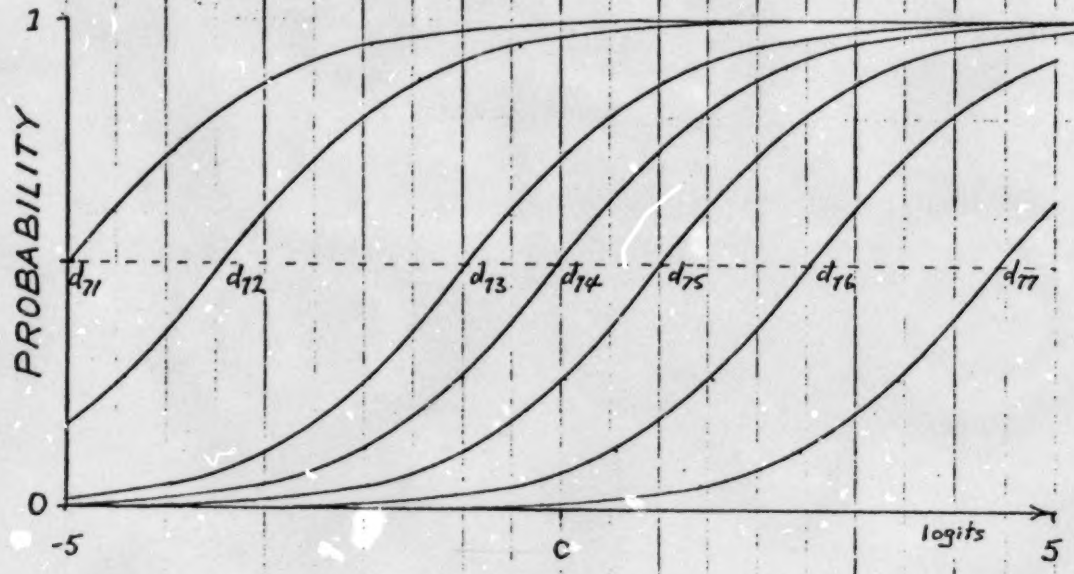
> "<u>Overall Competence:</u> This final rating is concerned with your judgment of the individual's overall performance in the clinical setting taking into account all of the ratings you have made above".

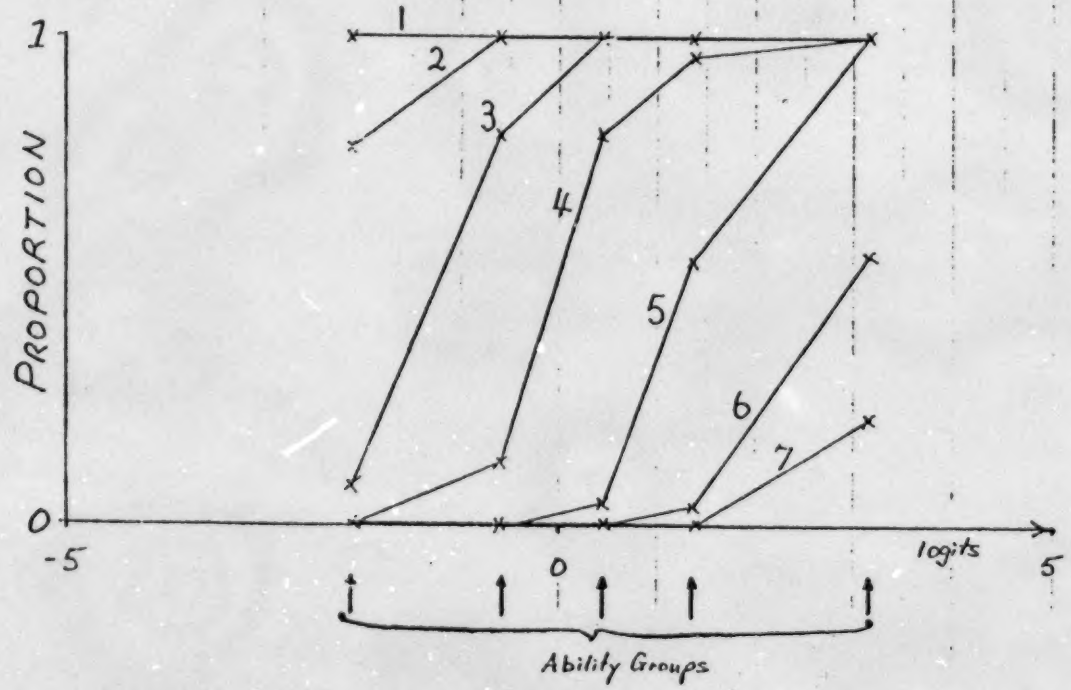
This instruction to raters to base their ratings on Item 7 on their other six ratings violates the local independence assumption of the Partial Credit model

FIGURE 5

MODEL PROBABILITIES AND
OBSERVED PROPORTIONS

(ITEM 7)





(From Masters, 1980, p310)

and almost certainly accounts for the unexpectedly high item-total correlation for this item. Typically, very highly discriminating items are viewed as "better" items than those of average discrimination. Our experience suggests that high item discrimination can result from a violation of the local independence assumption. Ironically, rather than assisting in the definition of a coherent variable, items like Item 7 actually confound the definition of a variable.